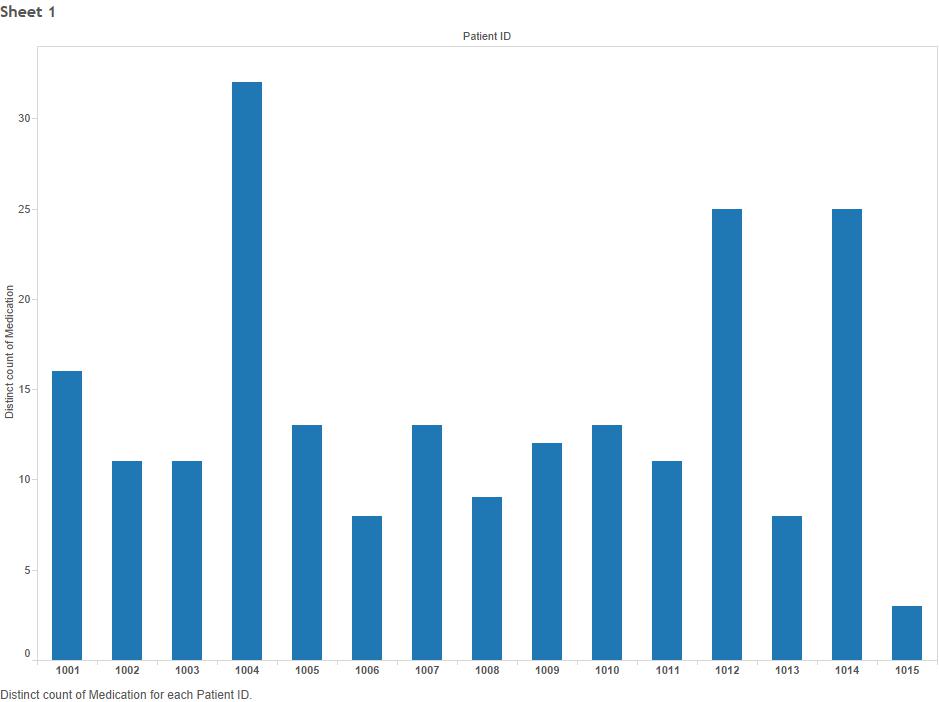
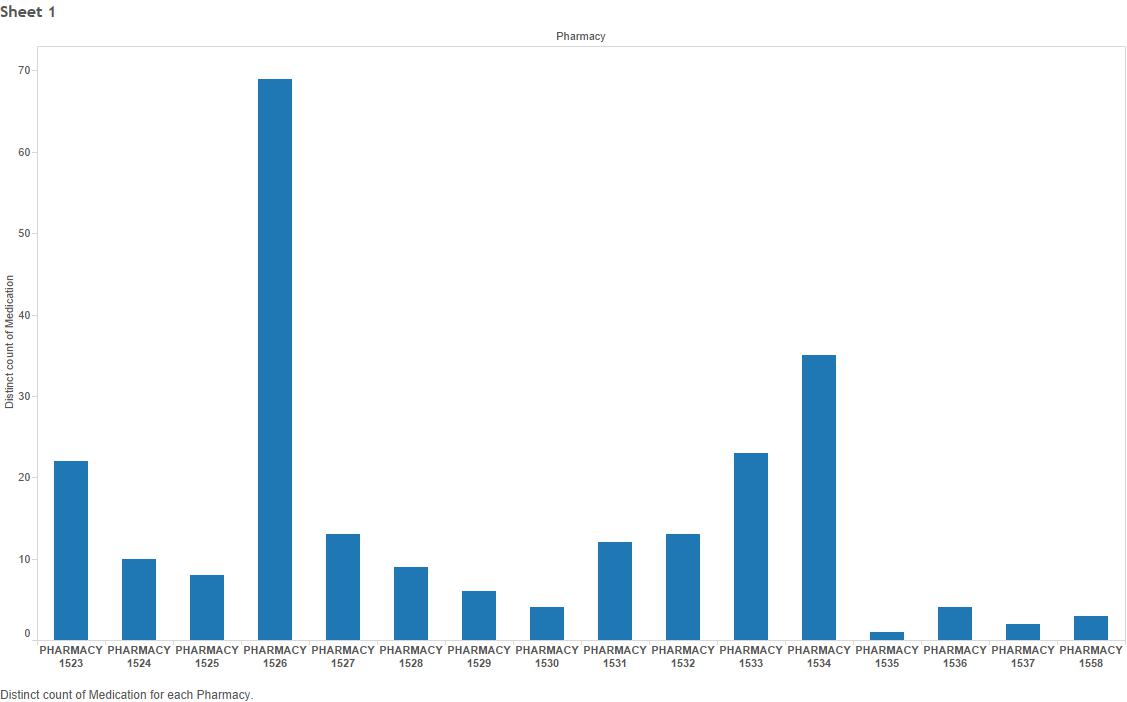
Understanding the Raw Data

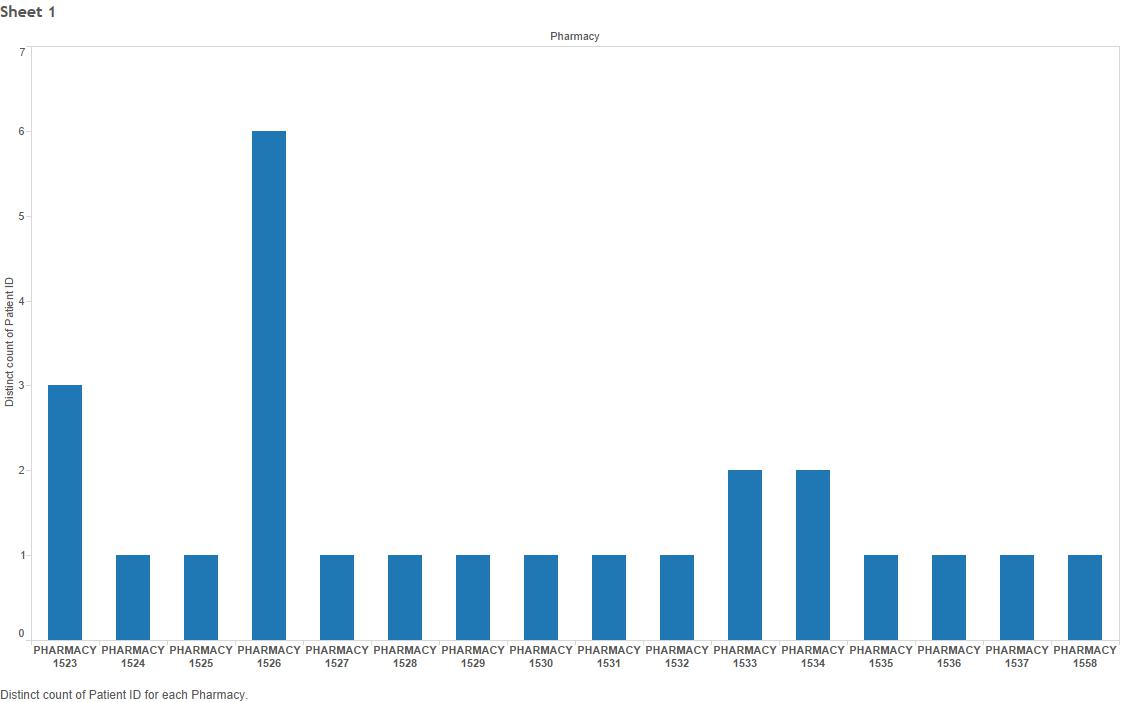
1. Patient vs No. of Medicines: The below plot shows the distribution of medicines between different patients. It shows that some patients are taking a large number of medicines than other medicines.



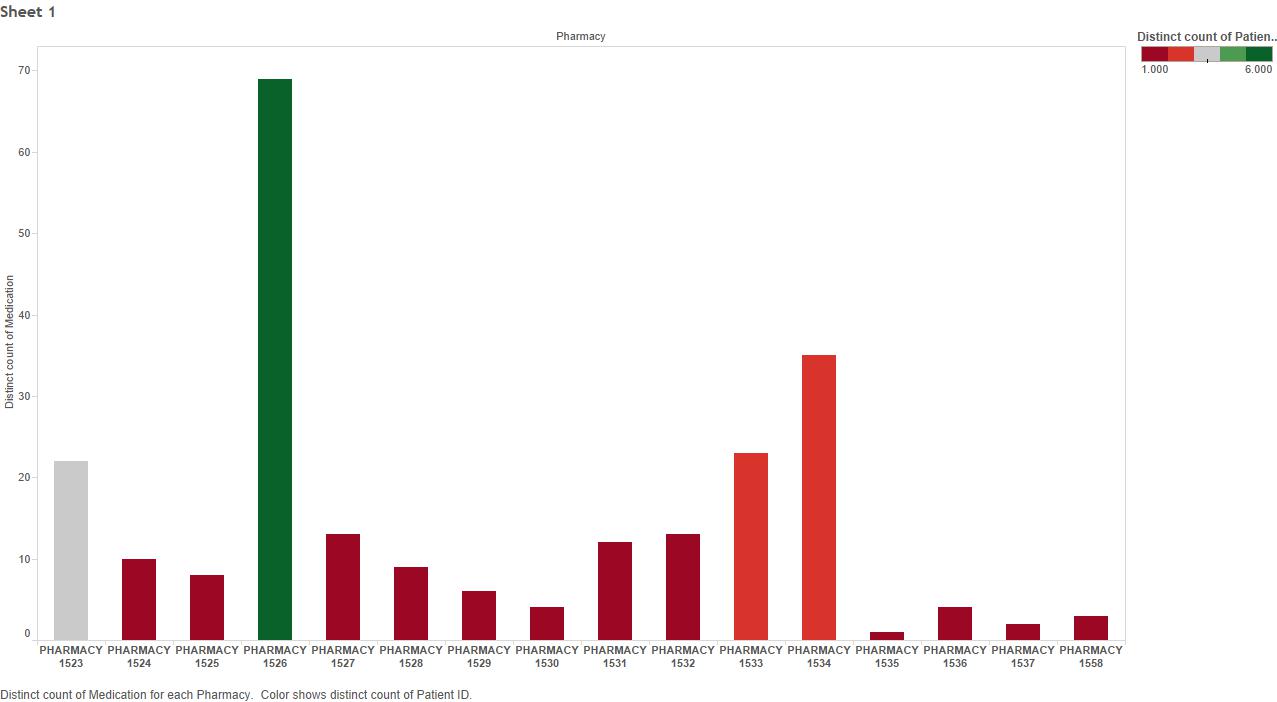
1. Distribution of medicines among pharmacies: The below plot shows that the medicine distribution among patients is varied. Only a few pharmacies are selling a variety of medications (at least in the data we have).



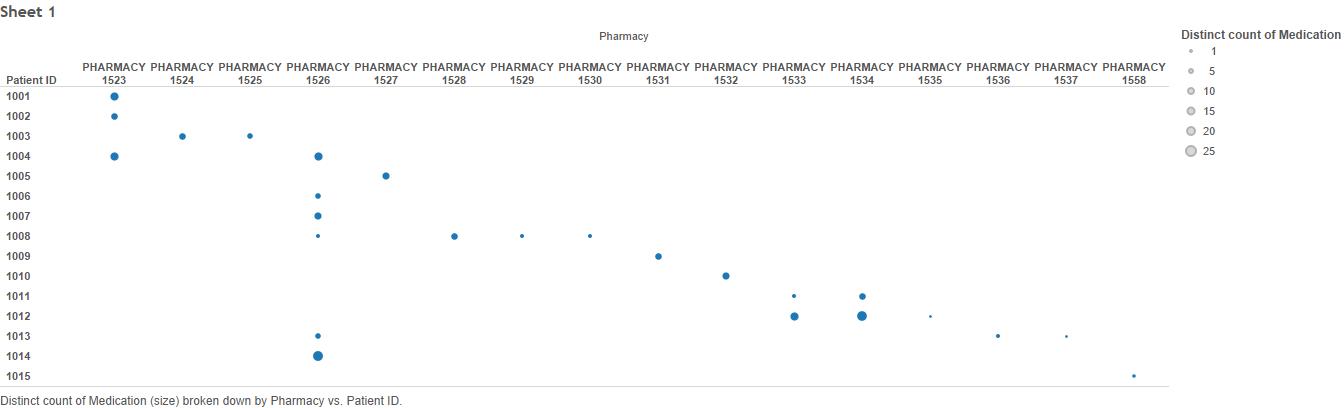
1. Distribution of patients among pharmacies: The below plot shows that the distribution of patients among pharmacies is very unique. At least in the data we have, it shows that most of the pharmacies serve a single patient and only a few pharmacies serve multiple patients.



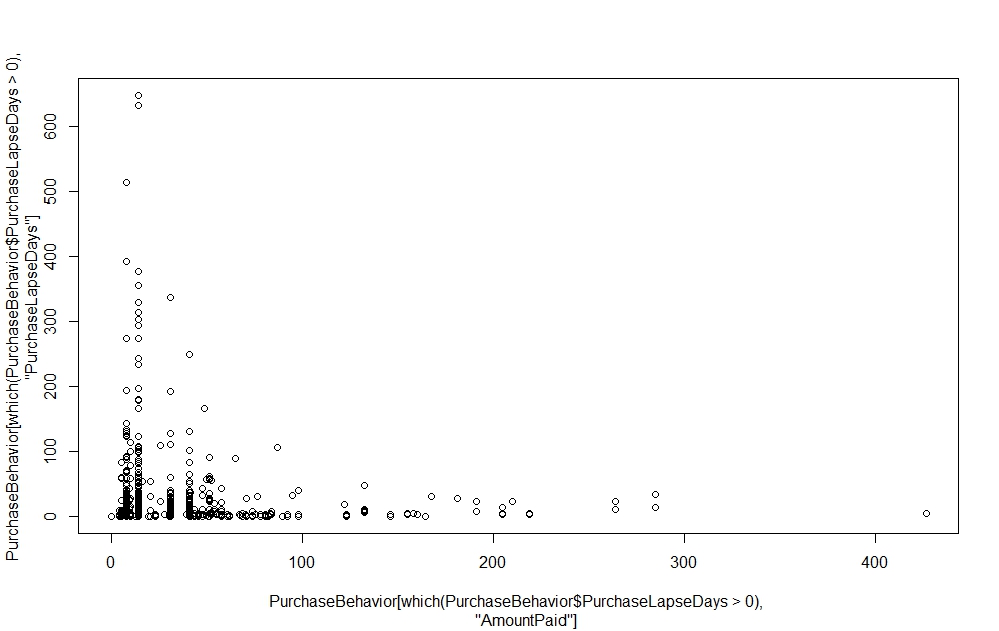
1. Pharmacy-Medicine vs No. of Patients: This is a combination of the previous two plots where the color is showing the number of medicines bought at each pharmacy. It simply implies that those pharmacies that serve multiple patients are selling more unique number of medicines.



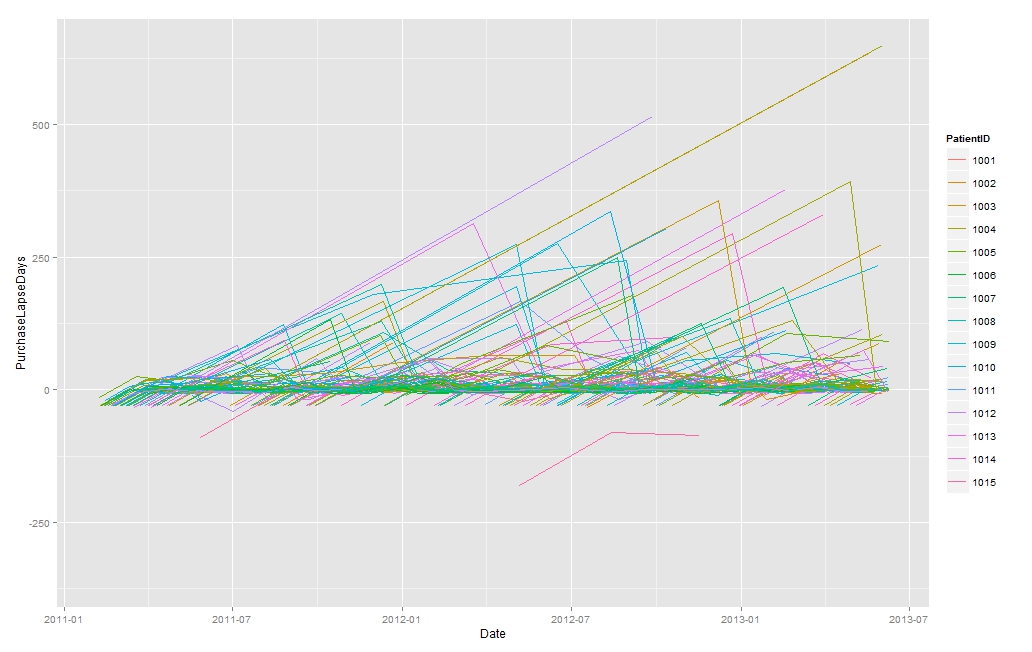
1. Pharmacy-Patient vs No. of Medicines: The below plot shows relatively how many medicines each patient has bought from each pharmacy. This shows that the combination of medicine and pharmacy plays a significant role.



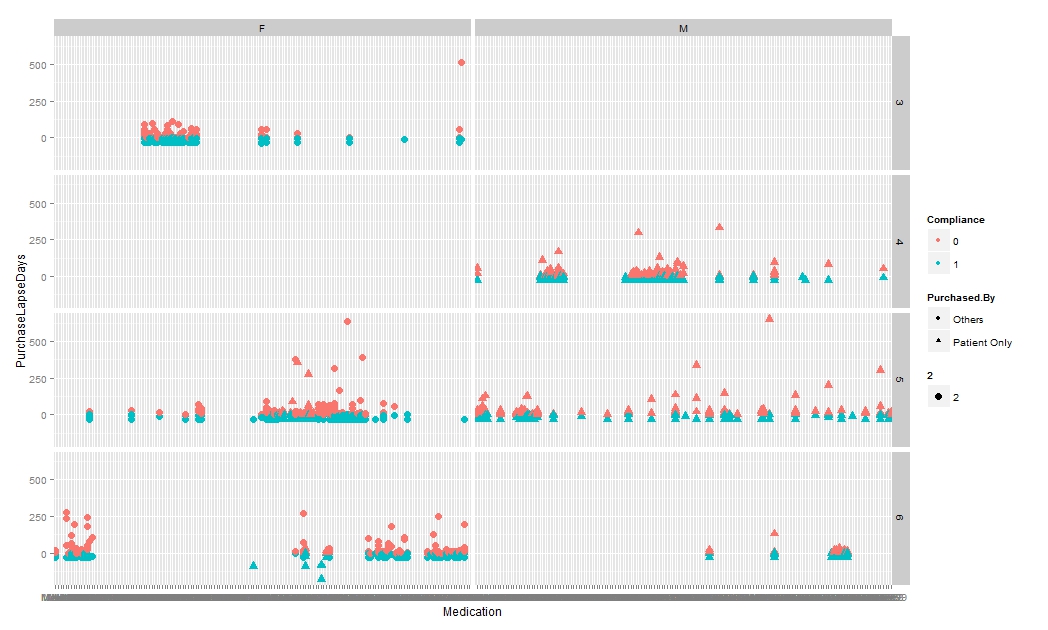
1. Purchase Lapse Days: This is the number of days lapsed by a patient in buying a medicine. For example, if a patient bought medicine for 30 days today and came back again after 40 days for refill, then the purchase lapse days are 10 days (40 – 30 = 10).
2. Purchase Lapse Days vs Amount Paid: The below plot shows how the purchase lapse days for all patient-medicine combinations is varying with the amount paid at each purchase. This shows that the patients are lapsing more with medicines that cost relatively less.

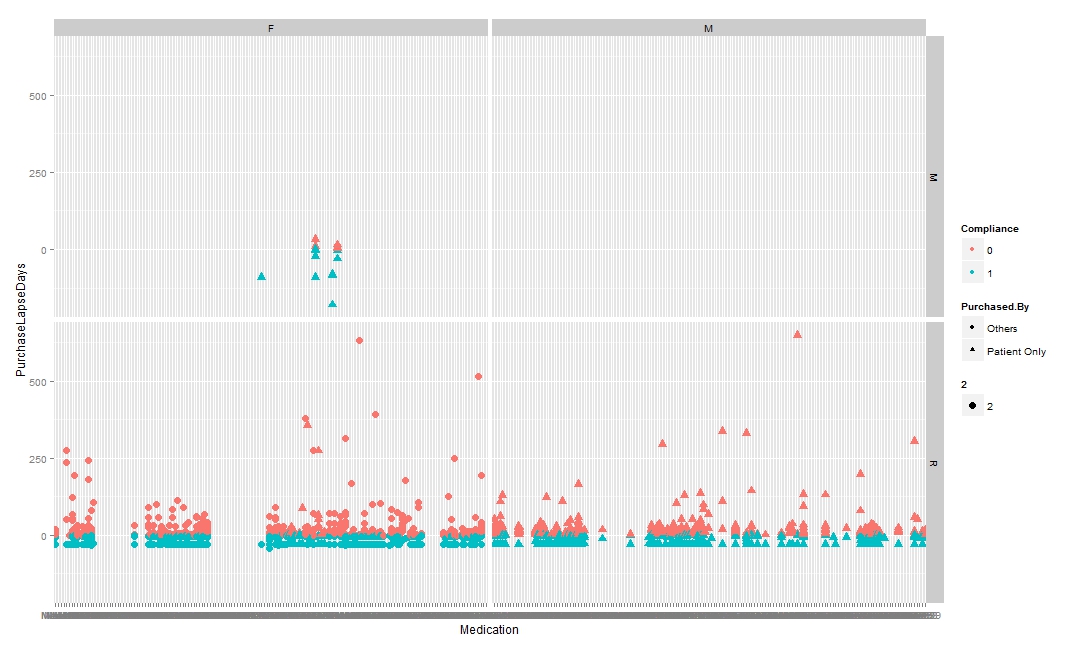


1. Purchase Lapse Days vs Date: The below plot shows how the purchase lapse days for each patient is varying with each purchase. It can be noticed that most of the patients are not lapsing by a huge margin (approximately 30 days). However there are a few that are lapsing by a very huge margin (peaks in the graph) and then falling back in order. Based on this we have made an assumption that when such a peak is observed, let us take that as a new round of purchase and reset the purchase lapse days to 0 from that purchase.



1. Data distribution by Age group, gender and purchase method (Retail Store or Mail / through self or others): The below set of graphs plot the data with multiple dimensions which give the below insights.
   1. All men bought their own medicines.
   2. Very few women bought their own medicines.
   3. Some medicines are particular to age groups and gender.
   4. The purchase by mail was availed by only women (digging into data showed that this was only one woman). Therefore this column of data is insignificant.



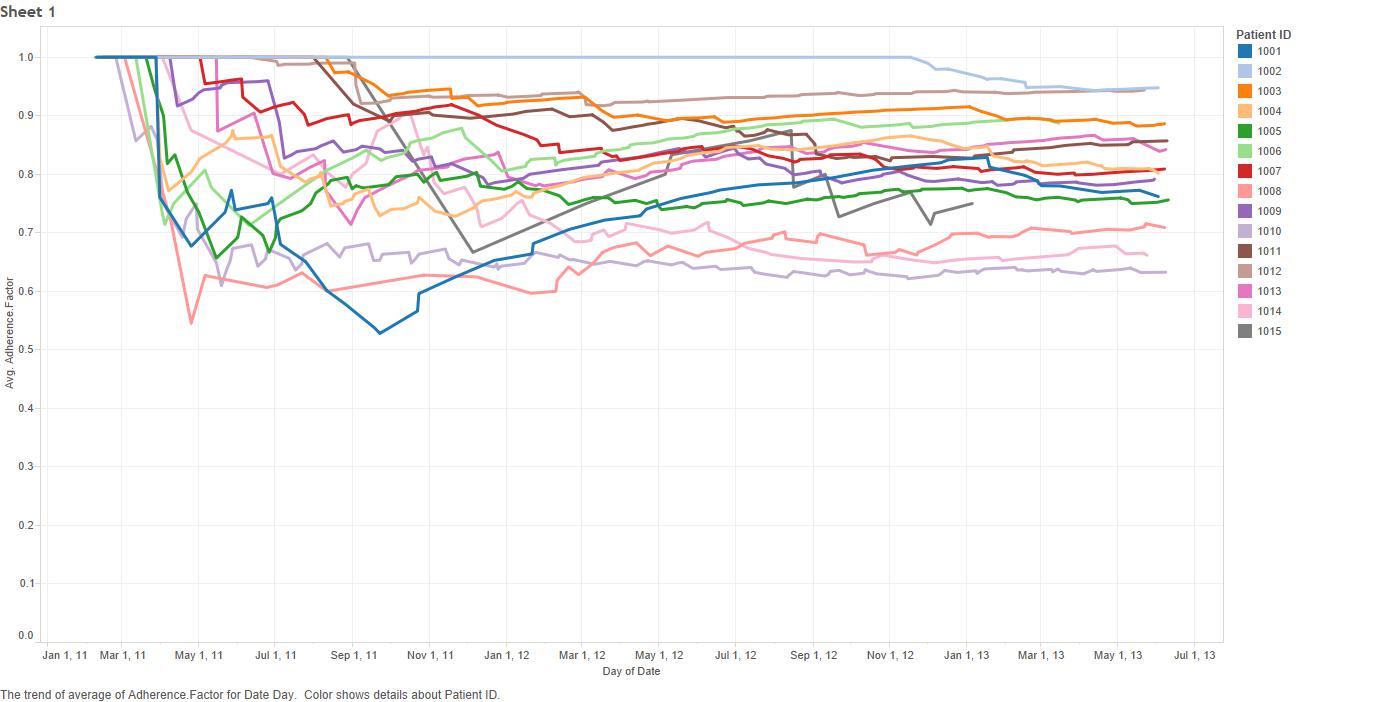


Feature Engineering the Data after Understanding:

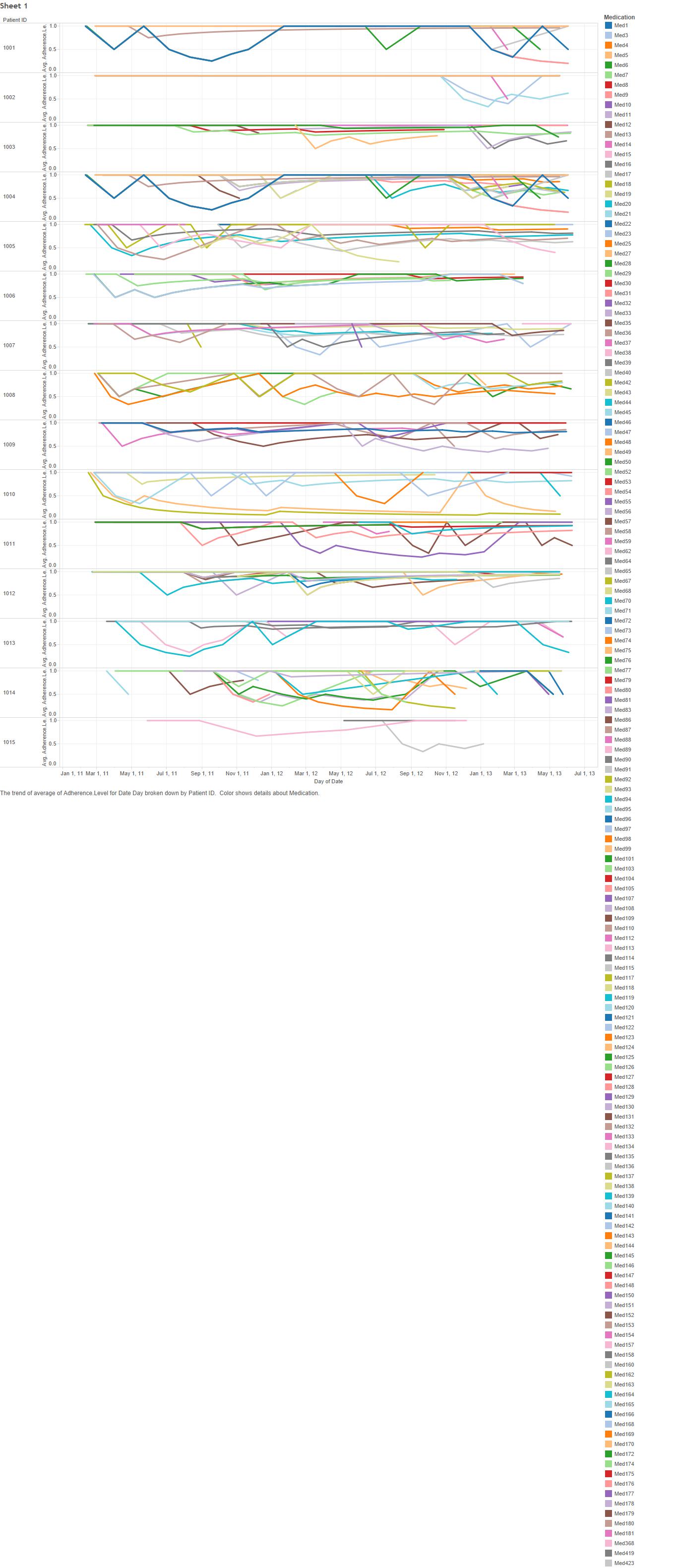
With all the above insights, the below assumptions were made.

1. If the purchase lapse days cross a certain limit (for example 4 days), let us consider that purchase to be non-adherent.
2. If the purchase lapse days cross a certain large limit (for example 21 days), let us consider that purchase to be a fresh purchase rather than non-adherence.
3. Let us calculate a value, Lapse.Factor, which will identify the change in consecutive Purchase Lapse Days both in direction and magnitude. For example, if a patient lapses by 20 days this time and then by 10 days next time, then the Lapse.Factor will be -10 (10 -20 = -10).
4. Let us calculate a value, Adherence.Level, for a patient-medicine combination of purchase record which gives the adherence percentage of that patient for that medicine.
5. Let us calculate a value, Adherence.Factor, for a patient which gives the total adherence percentage of that patient till that date.

The below plot shows the behavior of Adherence.Factor for each patient for all the data that we have. Clearly, this has a pattern where the adherence first decreases and then fluctuates.



The below plot shows the behavior of Adherence.Level for each patient-medicine combination that we have. It can be observed that the pattern is somewhat similar to Adherence.Level.



Model Building: We will now have to build a model that will identify the pattern we have observed in the feature engineered data with the insights we have observed in the raw data. After building a linear regression model considering Medication, Age, Sex, Purchased.By, AmountPaid, PurchaseLapseDays, Lapse.Factor, Adherence.Level, Adherence.Factor and treating Next.Lapse (the next purchase lapse days that we already know from the data) as the target variable, we get the below results.

Residuals:

Min 1Q Median 3Q Max

-18.602 -2.589 -0.226 2.089 18.921

This means that the model is able to predict the next purchase lapse days accurately where 68% of the result is within the error margin of about 2 days (1st Quadrant).

Running ANOVA (Analysis of Variance) on this model gives us the below significances to the independent variables.

Analysis of Variance Table

Response: Next.Lapse

Df Sum Sq Mean Sq F value Pr(>F)

Medication 130 13847.3 106.52 4.3978 < 2.2e-16 \*\*\*

Age 3 687.8 229.25 9.4651 3.504e-06 \*\*\*

Sex 1 71.8 71.80 2.9642 0.085385 .

Pharmacy 9 557.6 61.96 2.5580 0.006496 \*\*

AmountPaid 1 1.0 1.04 0.0429 0.835948

PurchaseLapseDays 1 544.0 544.03 22.4612 2.402e-06 \*\*\*

Lapse.Factor 1 110.9 110.91 4.5791 0.032566 \*

Adherence.Level 1 28.3 28.34 1.1700 0.279622

Adherence.Factor 1 219.9 219.89 9.0787 0.002640 \*\*

Adherence.Class 1 7.2 7.25 0.2992 0.584503

Residuals 1194 28919.7 24.22

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1